Homework 7

Anthony Zhang - SDS 315 UT Austin, az22589, https://github.com/antzha630/HW-7

Contents

Github Link for R Script

---

1. Armfolding

Part A

The number of male students in the dataset is 106 while the number of female students in the dataset is 111. The sample proportion of males who folded their left arm on top is 0.4717 while the sample proportion of females who folded their left arm on top is 0.4234.

Part B

The observed difference in proportions between the two groups (males minus females) is 0.0483. It represents how much likely or less likely males were to fold their left arm on top compared to females solely based on the sample.

Part C

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c out of ca["Male", "1"] out of sum(armfold$Sex == "Male")a["Female", "1"] out of sum(armfold$
## X-squared = 0.51118, df = 1, p-value = 0.4746
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.08393731  0.18048668
## sample estimates:
##    prop 1    prop 2
## 0.4716981 0.4234234
```

The 95% confidence interval for the difference in proportions (males minus females) is -0.0838 to 0.1805. To find the confidence interval using R's built in function, I used prop.test. To find the confidence interval using formulas, I used the one below to find the standard error. Standard Error = sqrt((p_1 (1 - p_1) / n_1) + (p_2 (1 - p_2) / n_2)) In our problem, p_1 would be the sample proportion of males who fold their left arm on top. n_1 would be the number of males while p_2 would be the sample proportion of males who fold their right arm on top. n_2 is the number of females. These values were written in part A. After plugging in all the values, I got the standard error value as 0.0675. Then, I multipled the standard error to the critical value of 1.96. This crit value comes from the standard normal Z distribution and is used when trying to make a 95% confidence interval for population proportion. The margin of error I got was 0.1322 in which I added and subtracted it from the observed difference in proportions between the two groups to obtain the 95% confidence interval for the difference in proportions.

Part D

If we were to repeat this arm folding experiemnt many times, then we would expect that approximately 95% of the resulting confidence intervals would contain the true difference in population proportion of males and females who fold their left arm on top.

Part E

In my own words, standard error is the standard deviation of the sampling distribution of the difference in sampling proportions between men and females who fold their left hand on top. It shows how much the observed difference in proportions might vary across repeated random samples from the population. As a result, a smaller SE would show that the estimate is much more precise while a larger SE would show that the estimate is not as precise.

Part F

The sampling distribution is the distribution for the difference in sample proportions of male students who put their left hand on top minus female students who put their left hand on top over many hypothetical random samples from the Australian university's undergraduate population. The true population proportions or the actual proportions of male students or female students that put their left hand on top stays fixed. As a result, the difference between these proportions will also be fixed. On the other hand, the observed proportions of male students or female students that put their left hand on top varies from sample to sample. The difference between these proportions will not be fixed.

Part G

The central limit theorem is a mathematical theorem that justifies using the normal distribution to approximate the sampling distribution of the difference in sampling proportions. It says that if the sample size is large enough, the sampling distribution of the difference in two proportions is approximately normal even if the distribution of the original data isn't. In our case, the sample sizes of male and female are pretty big with 106 people and 111 people. Also, the number of successes and failures (left hand over, right hand over) for each group is pretty big for both the male and female groups.

Part H

Because the confidence interval contains 0, there isn't enough evidence to confidently conclude there is a sex difference in arm folding for the population between the male and female groups. Since most of the interval is above 0, there could be a sex difference in arm folding for te population between male and female groups and a bigger sample size could give clearer and statistical evidence that is needed to confidently conclude there is a sex difference in arm folding for the population between the male and female groups.

Part I

Sampling variability would cause the confidence intervals to be different across samples if this experiment was repeated many times with different random samples of university students. Essentially, the people in each sample will be different and the sample proportions for males and females that put their left hand on top will also differ too. Their confidence intervals will also be different too because of the unique MOEs, standard errors, sample proportions, and more. The collection of all those intervals would create a normal distribution that has the mean of the true population difference in proportions between males and females who put their left hand on top. 95% of the intervals would contain this true population difference in proportions.

2. Get out the vote

Part A

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c out of cb["1", "1"] out of sum(turnout$GOTV_call == 1)b["0", "1"] out of sum(turnout$GOTV_ca
## X-squared = 40.416, df = 1, p-value = 2.053e-10
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1432115 0.2638452
## sample estimates:
##    prop 1    prop 2
## 0.6477733 0.4442449
```

The proportion of those receiving a GOTV call who voted in 1998 is 0.6478 while the sample proportion of those not receiving a GOTV call who voted in 1988 is 0.4442. A large-sample 95% confidence interval for the difference interval for the difference in the proportions of voting in 1998 for those who received a GOTV call versus those who didn't is 0.1432 to 0.2638.

Part B

The variables voted1996, Age, and Majority are all confounding variables that prevent difference I observed in Part A from representing the true causal effect of the GOTV call on the likelihood that a person voted in 1998.

```
##        0        1
## 49.42534 58.30769
```

```
##        0        1
## 44.91404 55.41535
```

```
##
##  Welch Two Sample t-test
##
## data:  AGE by GOTV_call
## t = -6.9613, df = 256.33, p-value = 2.817e-11
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -11.395051  -6.369644
```

```
## sample estimates:
## mean in group 0 mean in group 1
##        49.42534        58.30769

##
##  Welch Two Sample t-test
##
## data:  AGE by voted1998
## t = -30.24, df = 10568, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -11.182008  -9.820602
## sample estimates:
## mean in group 0 mean in group 1
##        44.91404        55.41535
```

The mean age for people who received the GOTV call is 58.31 while the mean age for people who didn't recieve the GOTV call is 49.43. This means that people who recieved the GOTV call are, on average, older than those who didn't. The confidence intervals for the difference in means here do not include zero so there is a statistically significant difference. The mean age for people who voted in 1998 is 55.42 while the mean age for people who didn't vote in 1998 is 44.91. This means that people who voted in 1998 are, on average, older than those who didn't. The confidence intervals for the difference in means here do not include zero so there is a statistically significant difference.

```
##          0          1
## 0.01781818 0.02450798

##         0         1
## 0.3501818 0.4824855

##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of rowSumstab1 out of tab1 out of rowSums2 out of tab1
## X-squared = 4.1195, df = 1, p-value = 0.04239
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.107284916 -0.006443461
## sample estimates:
##    prop 1    prop 2
## 0.7447552 0.8016194

##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of rowSumstab2 out of tab2 out of rowSums2 out of tab2
## X-squared = 145.17, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.11746499 -0.08518083
## sample estimates:
##    prop 1    prop 2
## 0.7005697 0.8018926
```

The mean GOTV call rate for people registered with a major party is 0.0245 while the mean for people not registered with a major party is 0.0178. This means that people in a major party are, on average, more likely to receive a GOTV call than those who aren't. The confidence interval for the difference in proportions does not include zero, so there is a statistically significant difference. The mean voting rate in 1998 for people registered with a major party is 0.482 while the mean for those not registered with a major party is 0.350. This means that people in a major party are, on average, more likely to have voted in 1998 than those who aren't. The confidence interval for the difference in proportions does not include zero, so there is a statistically significant difference.

```
##          0          1
## 0.01409849 0.03038149
```

```
##         0         1
## 0.2293487 0.6397376
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of rowSumstab3 out of tab3 out of rowSums2 out of tab3
## X-squared = 32.047, df = 1, p-value = 1.505e-08
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2389791 -0.1245081
## sample estimates:
##    prop 1    prop 2
## 0.5308070 0.7125506
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of rowSumstab4 out of tab4 out of rowSums2 out of tab4
## X-squared = 1834.1, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.4297118 -0.3956805
## sample estimates:
##    prop 1    prop 2
## 0.3496984 0.7623946
```

The mean GOTV call rate for people who voted in 1996 is 0.0304 while the mean for people who did not vote in 1996 is 0.0141. This means that prior voters are, on average, more likely to receive a GOTV call than those who did not vote in 1996. The confidence interval for the difference in proportions does not include zero, so there is a statistically significant difference. The mean voting rate in 1998 for people who voted in 1996 is 0.640 while the mean for people who did not vote in 1996 is 0.229. This means that people who voted in 1996 are, on average, more likely to vote again in 1998. The confidence interval for the difference in proportions does not include zero, so there is a statistically significant difference.

Part C

```
##
## Call:
## matchit(formula = GOTV_call ~ AGE + MAJORPTY + voted1996, data = turnout,
```

```
##     ratio = 5)
##
## Summary of Balance for All Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance         0.0297        0.0226          0.5130     1.3026    0.1572
## AGE             58.3077       49.4253          0.4475     1.1228    0.1114
## MAJORPTY         0.8016        0.7448          0.1426          .    0.0569
## voted1996        0.7126        0.5308          0.4016          .    0.1817
##           eCDF Max
## distance    0.2499
## AGE         0.2229
## MAJORPTY    0.0569
## voted1996   0.1817
##
## Summary of Balance for Matched Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance         0.0297        0.0297          0.0001      1.004    0.0000
## AGE             58.3077       58.2664          0.0021      1.008    0.0006
## MAJORPTY         0.8016        0.8073         -0.0142          .    0.0057
## voted1996        0.7126        0.7126         -0.0000          .    0.0000
##           eCDF Max Std. Pair Dist.
## distance    0.0057          0.0001
## AGE         0.0057          0.0027
## MAJORPTY    0.0057          0.0183
## voted1996   0.0000          0.0000
##
## Sample Sizes:
##           Control Treated
## All         10582     247
## Matched      1235     247
## Unmatched    9347       0
## Discarded       0       0
```

This output shows that the matching process reduced the control group sample size in order to get a better balance and to try to account for the confounding variables. The program discarded 9,347 unmatched control units and matched 247 treated units to 1,235 matched control units. None of the treated units were thrown away.

```
##        0        1
## 58.26640 58.30769
```

```
##
##  Welch Two Sample t-test
##
## data:  AGE by GOTV_call
## t = -0.02987, df = 350.55, p-value = 0.9762
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.760374  2.677783
## sample estimates:
## mean in group 0 mean in group 1
##        58.26640        58.30769
```

After using matching to create the new dataset and adjusting for the confounding variables (vote1996, age, majorpty), we can see that the mean age for people who received the GOTV call vs didn't (58.27 vs 58.31) was pretty close. This difference is much smaller compared to the difference before matching. The confidence interval containing zero shows that there is no significant difference.

```
##         0         1
## 0.8072874 0.8016194
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of rowSumstab_party out of tab_party out of rowSums2 out of tab_party
## X-squared = 0.042347, df = 1, p-value = 0.837
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04871171  0.06004775
## sample estimates:
##    prop 1    prop 2
## 0.8072874 0.8016194
```

We can see that the proportion of people registering to a major party versus proportion of people not registered to a major party (0.8016 vs 0.8073) was pretty close. This difference is much smaller compared to the difference before matching. The confidence interval contains zero which shows there is no significant difference.

```
##         0         1
## 0.7125506 0.7125506
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of rowSumstab_vote96 out of tab_vote96 out of rowSums2 out of tab_vote96
## X-squared = 2.6633e-29, df = 1, p-value = 1
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.06182709  0.06182709
## sample estimates:
##    prop 1    prop 2
## 0.7125506 0.7125506
```

We can see that the proportion of people who voted in 1996 versus proportion of people who didn't vote in 1996 (0.7126 vs 07126) was exactly the same. This difference is much smaller compared to the difference before matching. The confidence interval contains zero which shows there is no significant difference.

```
## [1] 0.6477733
```

```
## [1] 0.5692308
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c out of csum(turnout_matched$GOTV_call == 1 & turnout_matched$voted1998 == 1) out of gotv_yes
```

```
## X-squared = 5.2206, df = 1, p-value = 0.02232
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.01288268 0.14420234
## sample estimates:
##    prop 1    prop 2
## 0.6477733 0.5692308
```

For the matched dataset: The proportion of those receiving a GOTV call who voted in 1998 is 0.6478 while the sample proportion of those not receiving a GOTV call who voted in 1988 is 0.5692. A large-sample 95% confidence interval for the difference interval for the difference in the proportions of voting in 1998 for those who received a GOTV call versus those who didn't is 0.0129 to 0.1442. Since the confidence interval doesn't contain 0, we conclude that receiving a GOTV call had a statistically significant positive effect on the likelihood of voting in the 1998 election. People who recieved a GOTV call were about 7.85% more likely to vote than people who didn't recieve a GOTV call on average. We can make this conclusion because we controlled for the confounding variables through matching.
